

Semi-analytical model of the contact resistance in two-dimensional semiconductors

Roberto Grassi,^{1,*} Yanqing Wu,² Steven J. Koester,¹ and Tony Low¹

¹*Department of Electrical and Computer Engineering,
University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455, USA*

²*Wuhan National High Magnetic Field Center and School of Optical and Electronic Information,
Huazhong University of Science and Technology, Wuhan 430074, China*

(Dated: January 27, 2017)

Contact resistance is a severe performance bottleneck for electronic devices based on two-dimensional layered (2D) semiconductors, whose contacts are Schottky rather than Ohmic. Although there is general consensus that the injection mechanism changes from thermionic to tunneling with gate biasing, existing models tend to oversimplify the transport problem, by neglecting the 2D transport nature and the modulation of the Schottky barrier height, the latter being of particular importance in back-gated devices. In this work, we develop a semi-analytical model based on Bardeen's transfer Hamiltonian approach to describe both effects. Remarkably, our model is able to reproduce several experimental observations of a metallic behavior in the contact resistance, i.e., a decreasing resistance with decreasing temperature, occurring at high gate voltage.

Introduction— 2D layered semiconducting materials, such as transition metal dichalcogenides (TMDs) and black phosphorus (BP), have many interesting electrical and optical properties [1–7], but tend to form Schottky barriers (SB) at the interfaces with metal contacts, resulting in a large contact resistance that severely degrades the device performance [8–10].

Thermionic emission [11] is commonly assumed when extracting the SB height from temperature-dependent current measurements of field-effect transistors (FETs). In [12], it was pointed out that this procedure is correct only at the gate voltage corresponding to the flat-band condition. Considering an n-type device, for example, above the flat-band voltage, the conduction band edge in the channel is higher than at the interface with the contact, hence electrons traversing the channel see a larger barrier than the SB height. Below the flat-band voltage, tunneling starts to contribute and the thermionic emission theory loses validity. As a result, this can lead to unphysical negative SB heights [13, 14]. Furthermore, experiments show that, as opposed to the insulating behavior of a SB contact, the two-terminal resistance [15] as well as contact resistance [16] can decrease with decreasing temperature at high gate voltage. The origin of this metallic behavior is debated and not yet clarified [12, 17].

Recently, a model for SB FETs has been proposed in [18] and applied to extract the SB height and bandgap of BP devices. This model assumes one-dimensional transport and a bias-independent SB height. However, in a typical geometry with a top contact to a multilayer 2D semiconductor as in the sketch of Fig. 1a, transport is inherently 2D. To be precise, due to quantization, the SB height Φ_1 to a 2D semiconductor should be defined as the difference between the edge E_1 of the first energy subband in the semiconductor and the Fermi level

of the metal μ (see the schematic band profile in Fig. 1b for an n-type device). Transport occurring at energies above (below) E_1 is generally referred to as “thermionic” (“tunneling”). However, in the presence of a back gate, the subband edge and thus the SB height are expected to be modulated by the vertical electric field. When E_1 is lower than the bulk band edge at the interface with the metal, even the electrons traversing the junction at energies above E_1 see a tunneling barrier in the vertical (i.e., z) direction and a new transport regime arises.

In this paper, we present a semi-analytical model of the contact resistance to multilayer 2D semiconductors in this “vertical tunneling” regime. The model is based on a triangular barrier approximation of the vertical potential profile in the semiconductor underneath the contact. 2D transport is separated into a sequence of two 1D mechanisms: (i) quantum tunneling through the SB at the metal-to-semiconductor interface, followed by (ii) semi-classical “diffusive” transport across the semiconductor (the source of scattering being the in- and out-tunneling across the SB). The model is benchmarked against numerical solutions of the 2D quantum transport problem and employed to study the dependence of the contact resistance on vertical electric field and temperature. We show that, when the SB height is sufficiently lowered by the vertical electric field, contact resistance shows a metallic behavior with temperature, as observed in experiments. The model predicts a smooth transition from a thermionic-like regime at low electric field, where the tunneling barrier is almost transparent, to a true vertical tunneling regime at high electric field. In the former case, the extraction method of the SB height based on thermionic emission theory can still be applied.

Model— We consider a single planar junction of length L_x between a metal and a multi-layer 2D semiconductor with thickness a (Fig. 1a). A vertical electric field is created inside the semiconductor by the presence of a back gate. Let x and z be the longitudinal and vertical directions, respectively. The device is uniform in the y direction. We focus on the portion of the semiconductor

*Electronic address: rgrassi@umn.edu

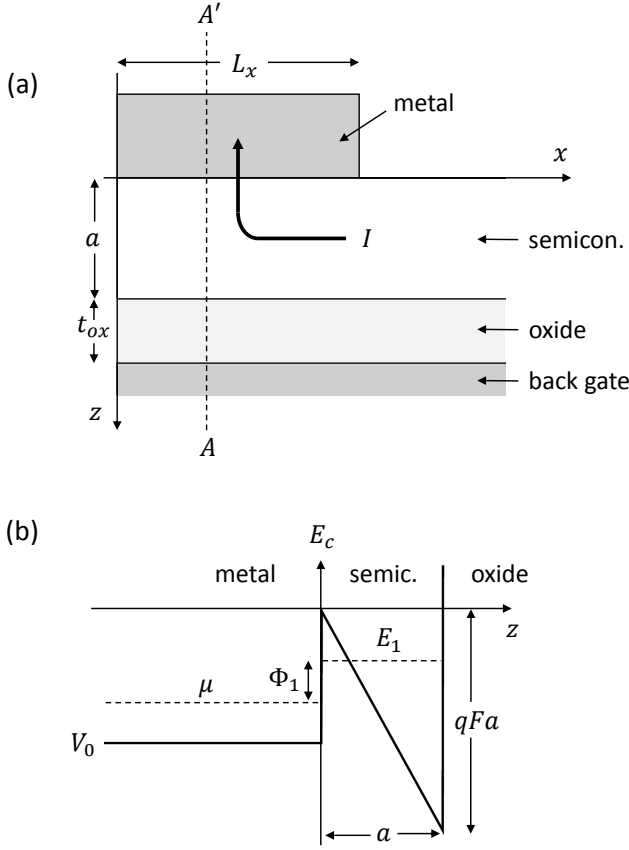


FIG. 1: (a) Cross-section of the device structure under consideration. The path of the current flow is also indicated schematically. (b) Triangular barrier model of the conduction band edge profile along the cut line A–A' in (a). The top of the barrier is taken as the energy reference. F is the vertical electric field in the semiconductor. The SB height Φ_1 is the energy difference between the first subband edge E_1 in the semiconductor and the Fermi level of the metal μ .

covered by the metal (contact region) assuming that the uncovered part (channel) simply acts as a “reflectionless” contact or semi-infinite lead [19]. A current can flow as indicated in Fig. 1a.

We neglect hole transport and discuss only injection of electrons from the metal to the conduction band of the semiconductor. A simple single-valley effective mass Hamiltonian, with the same values of the effective masses $m_{x,y,z}$ in the metal and in the semiconductor, is adopted. Such effective mass model has been shown to provide an accurate description of the out-of-plane quantization in multilayer BP [20]. For the case of multilayer TMDs, one should employ a more fundamental tight-binding model [21]. Whereas the form of the equations will be different, the general trends are not expected to change significantly. The bulk band edge profile is assumed to be uniform along the x direction and is approximated with a triangular barrier along the z direction, as shown in Fig. 1b, where the value of the band edge in the metal V_0 is chosen low enough so that it provides significant

density of states at the Fermi level. Within this non-self-consistent approximation, which is valid at low carrier concentration, the magnitude of the vertical electric field F in the semiconductor is simply proportional to the voltage V_G applied between the back gate and the top metal:

$$F = \frac{V_G}{a + (\epsilon_s/\epsilon_{ox})t_{ox}}, \quad (1)$$

where the workfunctions of the two metals are taken to be equal, t_{ox} is the back oxide thickness, and ϵ_s and ϵ_{ox} are the dielectric constants of the semiconductor and oxide, respectively.

Due to vertical confinement, the energy spectrum in the semiconductor splits into a set of discrete 2D subbands. Within a triangular well approximation, the subband edges E_i (i positive integer) can be computed as [22]

$$E_i = -qF \left(a - \frac{|\zeta_i|}{k_F} \right), \quad (2)$$

where the energy reference is taken at the top of the barrier in Fig. 1b, q is the elementary electric charge, the wavevector k_F is defined as

$$k_F = \left(\frac{2m_z q F}{\hbar^2} \right)^{1/3}, \quad (3)$$

and ζ_i are the zeros of Airy's function, i.e., $\text{Ai}(\zeta_i) = 0$, which can be approximated as [23]

$$\zeta_i \approx - \left[\frac{3\pi}{8} (4i - 1) \right]^{2/3}. \quad (4)$$

We limit the discussion to the case $E_i < 0$. Indeed, the subband description loses validity above the barrier.

We assume that transport within the semiconductor can be described by a set of decoupled 1D Boltzmann's transport equations [24], one for each subband, where the tunneling from the metal to the semiconductor and vice versa is included as a scattering mechanism. The corresponding relaxation time τ_i , or inverse of the probability rate that an electron originally in the k -space state (k_x, k_y) of the i -th subband tunnels into the metal, is computed according to Bardeen's transfer Hamiltonian theory [25–27], which has been recently applied to describe tunneling in vertical heterostructures of 2D materials [28, 29] and electron-hole bilayer tunnel FETs [30, 31]. In the limit of large L_x , we get

$$\frac{1}{\tau_i} = \begin{cases} \frac{1}{h} \frac{\hbar^2}{2m_z} \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} \frac{4\sqrt{-E_i(E_i - V_0)}}{-V_0} e^{-2\gamma_0}, & V_0 < E_i < 0 \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where h is Planck's constant, $\hbar = h/(2\pi)$, γ_0 is defined as

$$\gamma_0 = \frac{2}{3} \zeta_0^{3/2}, \quad (6)$$

$$\zeta_0 = -k_F \frac{E_i}{qF} = k_F a - |\zeta_i|, \quad (7)$$

and $\text{Ai}'(\zeta_i)$ is the derivative of Airy's function evaluated at ζ_i , which can be approximated as [23]

$$|\text{Ai}'(\zeta_i)| \approx \frac{1}{\sqrt{\pi}} \left[\frac{3\pi}{8}(4i-1) \right]^{1/6}. \quad (8)$$

Note that τ_i is independent of both k_x and k_y . The tunneling current is computed from the x -dependent distribution function of each subband, which is obtained by solving Boltzmann's transport equation with τ_i as the scattering relaxation time and with appropriate boundary conditions. In particular, we assume that the electrons are backscattered at the left end of the contact region. Differentiating the tunneling current with respect to the applied bias V_D and taking the limit $V_D \ll k_B T/q$ (k_B is Boltzmann's constant and T the temperature) gives us the low bias conductance G or inverse of contact resistance (per unit width). We obtain the semi-analytical expression

$$G = \frac{2q^2}{h} \int_{-\infty}^{\infty} d\varepsilon \bar{T}(\varepsilon) \left(-\frac{\partial F_0}{\partial \varepsilon} \right), \quad (9)$$

$$F_0(\varepsilon) = \sqrt{\frac{m_y k_B T}{2\pi \hbar^2}} \mathcal{F}_{-1/2} \left(\frac{\mu - \varepsilon}{k_B T} \right), \quad (10)$$

where ε is the total energy for electrons with $k_y = 0$ and $\mathcal{F}_{-1/2}$ the Fermi-Dirac integral of order $-1/2$. The total transmission function \bar{T} is defined as

$$\bar{T}(\varepsilon) = \sum_i T_i(\varepsilon), \quad (11)$$

with the transmission probability T_i of each subband given by

$$T_i(\varepsilon) = \begin{cases} 0, & \varepsilon < E_i \\ 1 - e^{-\frac{2L_x}{\lambda_i}}, & \varepsilon > E_i \end{cases}, \quad (12)$$

where $\lambda_i = |v_x| \tau_i$ is the mean free path related to tunneling and $|v_x| = \sqrt{2(\varepsilon - E_i)/m_x}$ is the longitudinal carrier velocity. In the Supporting information, we provide a detailed derivation of the model.

Results— In Fig. 2 we plot the tunneling rate $1/\tau_i$ of the first two subbands as a function of electric field and semiconductor thickness. The predicted tunneling rate goes to zero at small F or a because the Bardeen model does not account for the above-the-barrier regime at $E_i > 0$. At high electric field or large semiconductor thickness, the exponential term in (5) is dominating. In this regime, an increase of F or a results in a decrease of the scattering rate $1/\tau_i$. This can be understood by noting that, since both k_x and k_y are conserved in the tunneling process, an electron can tunnel from the metal to the semiconductor only if its vertical energy is equal to E_i . However, according to (2), E_i shifts to lower energies with increasing F or a . Because of that shift, the tunneling distance, which is equal to $|E_i|/(qF)$ at the vertical energy E_i , becomes longer as F or a increase.

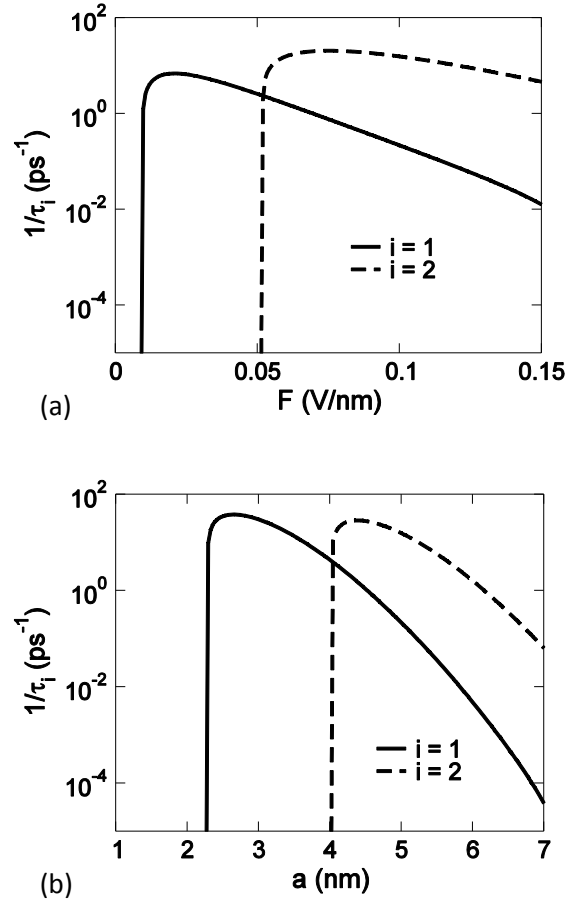


FIG. 2: Tunneling rate of the first ($i = 1$) and second ($i = 2$) subband, computed from (5) and plotted as a function of electric field F at fixed semiconductor thickness $a = 5$ nm in (a), and as a function of a at fixed $F = 0.1$ V/nm in (b). The other parameter values are: $m_z = 0.4m_0$ (m_0 is the free electron rest mass) and $V_0 = -0.5$ eV.

In order to benchmark the proposed model, we solve numerically the 2D Schrödinger equation with open boundary conditions using the Green function (GF) method [19] and assuming the same non-self-consistent triangular potential profile as in Fig. 1b. Details on the GF calculation can be found in the Supporting information. Fig. 3 shows the plot of the total transmission function $\bar{T}(\varepsilon)$ computed with the analytical expression in (11)–(12) and with GF for different sets of parameter values. The two models are in good general agreement. The transmission function increases by one at each energy corresponding to a subband edge E_i , indicating a resonant tunneling regime, and shows a decaying behavior between two successive subband edges. Indeed, different energies correspond to different k_x states. Since the length of the contact L_x is finite and the transfer length or average distance traveled by an electron in the semiconductor before tunneling into the metal is equal to the mean free path $\lambda_i = |v_x| \tau_i$, the probability of escaping into the metal is larger for the states closer to the

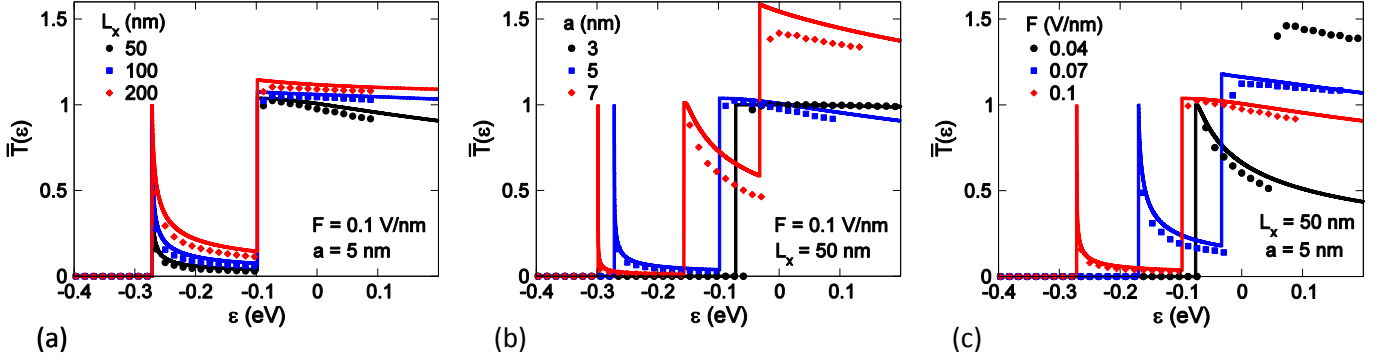


FIG. 3: Comparison between the transmission function vs. energy from the model in (12) (lines) and from GF (symbols) for different values of (a) contact length L_x , (b) semiconductor thickness a , and (c) electric field F . Other parameter values are: $m_x = 0.2m_0$, $m_z = 0.4m_0$, $V_0 = -0.5$ eV.

subband edge which have smaller velocity. As shown in Fig. 3, increasing the contact length tends to raise the transmission probability of each single subband to unity because the ratio L_x/λ_i between the contact length and the average distance before tunneling increases, which means that the electrons have more chances to enter the contact. The shift of E_i to lower energy as a or F increase is clearly seen from the shift of the transmission peaks in Fig. 3b and c, respectively. This is accompanied by a narrowing of the peaks, which is related to the decrease of $1/\tau_i$ discussed above.

It should be noted that, with reference to the generic subband of index i , our model predicts no vertical tunneling contribution at energies below E_i (see Eq. 12). This process could be possible in the case of a realistic band bending between the channel and the contact region. However, since tunneling is decreasing exponentially with the tunneling distance, the effect would be concentrated at the contact edge. Therefore, if E_i is sufficiently close to the metal Fermi level μ , the contribution from energies below E_i (lateral tunneling) is negligible compared to energies above E_i (vertical tunneling) because of the large surface-to-edge ratio of the contact.

Fig. 4a plots G , obtained by numerically computing the energy integral in (9) and resolved for the first two subbands, as a function of F and temperature T . It can be seen that G is a non-monotonic function of the electric field. This is consistent with our previous observations: the transmission probability reduces with increasing F because the tunneling distance increases. Fig. 4a shows that, at high electric field before the second subband starts to contribute significantly, the derivative $\partial G/\partial T$ is negative, i.e., contact resistance decreases with decreasing temperature. This has to do with the factor $-\partial F_0/\partial \epsilon$, which is plotted in Fig. 4b. It can be proved that its derivative with respect to temperature changes sign at the energy $\epsilon_0 \approx \mu + 0.857k_B T$, which has only a weak temperature dependence ($\epsilon_0 = -0.243$ and -0.228 eV at $T = 100$ and 300 K, respectively) and appears as a crossover point in Fig. 4b. As the transmission

function shifts to lower energy with increasing F (compare the plots at $F = 0.08$ and 0.1 V/nm in Fig. 4b), more contribution to the integral in (9) comes from the energy range where $\partial(-\partial F_0/\partial \epsilon)/\partial T < 0$ and eventually leads to $\partial G/\partial T < 0$.

The model in (9) allows for an analytical solution in two limiting cases. In order to simplify the discussion, we assume that only the first subband contributes to transport. Note that the energy range relevant for transport goes from E_1 to few $k_B T$'s above E_1 or μ , whichever is maximum. If $L_x \gg \lambda_1$ in this energy range, it follows that $T_1(\epsilon) \approx 1$ (i.e., an almost transparent barrier) and (9) simplifies to

$$G \approx \frac{2q^2}{h} F_0(E_1). \quad (13)$$

If, in addition, $\Phi_1 = E_1 - \mu \gg k_B T$, then (13) further reduces to

$$G \propto T^{1/2} \exp\left(-\frac{\Phi_1}{k_B T}\right), \quad (14)$$

which is the expression of the thermionic emission theory for a 2D system [32]. The prefactor is $T^{1/2}$ instead of $T^{3/2}$ because we are considering the low bias limit $V_D \ll k_B T/q$. Expression (14) implies that the SB height Φ_1 can be extracted from the slope of $\ln(G/T^{1/2})$ vs $1/T$. On the other hand, if $L_x \ll \lambda_1$ in most of the energy window for transport, one can derive (see Supporting information)

$$G \approx \frac{2q^2}{h} \frac{\sqrt{m_x m_y}}{h} \frac{L_x}{\tau_1} f_0(E_1), \quad (15)$$

where $f_0(E) = \{\exp[(E - \mu)/(k_B T)] + 1\}^{-1}$ is the Fermi-Dirac function. Fig. 4c compares the Arrhenius plot of $G/T^{1/2}$ computed with the rigorous model in (9) and the approximated expressions in (13) and (15). E_1 is calculated according to (2) in all three cases. Similar Arrhenius plots are commonly used to extract the SB height in experiments [13, 14, 16, 33]. For the chosen

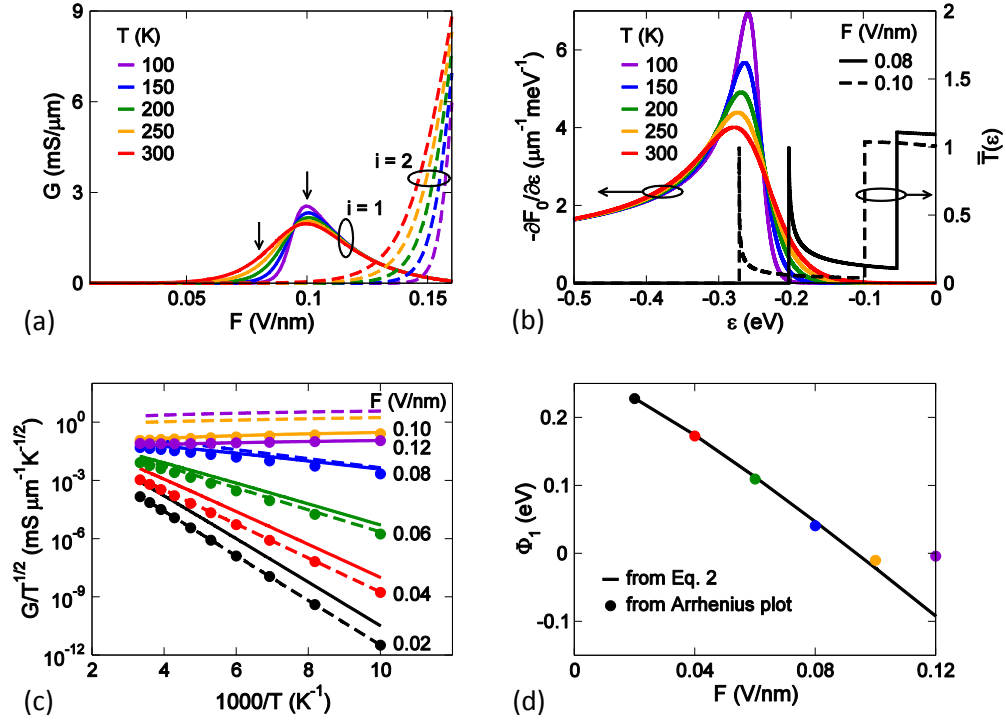


FIG. 4: (a) Conductance G vs. vertical electric field for different temperatures at $\mu = -0.25$ eV. The contributions of the first ($i = 1$) and second ($i = 2$) subband are separated. (b) Plot of $-\partial F_0/\partial \epsilon$ vs. energy for different temperatures at $\mu = -0.25$ eV. Superimposed are the spectra of the transmission function at the two electric field values indicated by arrows in (a). (c) Arrhenius plot of $G/T^{1/2}$ at different electric field values, computed from the rigorous model in (9) and the approximated expressions in (13) (dashed lines) and (15) (solid lines). (d) SB height Φ_1 vs. electric field, extracted from the average slope of the Arrhenius plot with symbols in (c) assuming thermionic emission (symbols), compared with the actual values from (2) (line). Other parameters values are: $m_x = 0.2m_0$, $m_y = m_0$, $m_z = 0.4m_0$, $V_0 = -0.5$ eV, $a = 5$ nm, $L_x = 50$ nm.

set of parameter values, approximation (13) is valid up to $F \approx 0.08$ V/nm. At higher electric field, the transmission function becomes increasingly peaked around the subband edge (see Fig. 4b) and (15) becomes a better approximation. A positive slope, or metallic behavior, is predicted at high electric field similar to what has been reported in experiments [13, 14]. In Fig. 4d, we plot the SB height obtained by fitting the data of the rigorous model in (9) with the thermionic expression (14), compared with the actual value of $\Phi_1 = E_1 - \mu$ from (2). It is seen that the extraction method based on the thermionic emission theory can provide good results at low electric field values, where the tunneling barrier is almost transparent. In the high-field regime, a fitting based on (15) would provide a more physical result.

We conclude by noting that the model presented in this work can be easily extended to account for a finite carrier mobility in the semiconductor by introducing an additional relaxation time τ_s (and a corresponding mean free path $\lambda_s = |v_x|\tau_s$) related to elastic scattering. The main

effect of scattering would be, for each subband, a shorter transfer length and a transmission probability that does saturate to unity in the limit of a long contact length. In the regime when only one subband is populated, the model could also be extended to include self-consistent electrostatics using the variational approach in [34].

Conclusions— In summary, we have demonstrated that the metallic behavior of the contact resistance observed in recent experiments can be explained by taking into account the modulation of the vertical tunneling due to the SB lowering with increasing electric field in back-gated devices. To the best of our knowledge, this transport regime has not been discussed before. The model also suggests a non-monotonic behavior of the contact resistance with respect to vertical electric field and semiconductor thickness. Our semi-analytical model provides a reasonable description of contact resistance in 2D semiconductors and could be useful for contact engineering in future 2D electronics.

- ica **102**, 10451 (2005).
- [2] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, *Nature nanotechnology* **7**, 699 (2012).
 - [3] X. Xu, W. Yao, D. Xiao, and T. F. Heinz, *Nature Physics* **10**, 343 (2014).
 - [4] G. Fiori, F. Bonaccorso, G. Iannaccone, T. Palacios, D. Neumaier, A. Seabaugh, S. K. Banerjee, and L. Colombo, *Nature nanotechnology* **9**, 768 (2014).
 - [5] T. Low, A. Chaves, J. D. Caldwell, A. Kumar, N. X. Fang, P. Avouris, T. F. Heinz, F. Guinea, L. Martin-Moreno, and F. Koppens, *Nature Materials* (2016).
 - [6] F. Koppens, T. Mueller, P. Avouris, A. Ferrari, M. Vitiello, and M. Polini, *Nature nanotechnology* **9**, 780 (2014).
 - [7] Z. Sun, A. Martinez, and F. Wang, *Nature Photonics* **10**, 227 (2016).
 - [8] S. Das and J. Appenzeller, *Nano letters* **13**, 3396 (2013).
 - [9] Y. Du, H. Liu, Y. Deng, and P. D. Ye, *ACS nano* **8**, 10035 (2014).
 - [10] N. Haratipour, M. C. Robbins, and S. J. Koester, *Electron Device Letters, IEEE* **36**, 411 (2015).
 - [11] S. M. Sze and K. K. Ng, *Physics of semiconductor devices* (John Wiley & sons, 2006).
 - [12] A. Allain, J. Kang, K. Banerjee, and A. Kis, *Nature Materials* **14**, 1195 (2015).
 - [13] L. Yu, Y.-H. Lee, X. Ling, E. J. Santos, Y. C. Shin, Y. Lin, M. Dubey, E. Kaxiras, J. Kong, H. Wang, et al., *Nano letters* **14**, 3055 (2014).
 - [14] A. Avsar, I. J. Vera-Marun, J. Y. Tan, K. Watanabe, T. Taniguchi, A. H. Castro Neto, and B. Ozyilmaz, *ACS nano* **9**, 4138 (2015).
 - [15] Y. Liu, H. Wu, H.-C. Cheng, S. Yang, E. Zhu, Q. He, M. Ding, D. Li, J. Guo, N. O. Weiss, et al., *Nano letters* **15**, 3030 (2015).
 - [16] X. Cui, G.-H. Lee, Y. D. Kim, G. Arefe, P. Y. Huang, C.-H. Lee, D. A. Chenet, X. Zhang, L. Wang, F. Ye, et al., *Nature nanotechnology* **10**, 534 (2015).
 - [17] B. Radisavljevic and A. Kis, *Nature materials* **12**, 815 (2013).
 - [18] A. V. Penumatcha, R. B. Salazar, and J. Appenzeller, *Nature communications* **6** (2015).
 - [19] S. Datta, *Electronic transport in mesoscopic systems* (Cambridge university press, 1997).
 - [20] G. Zhang, S. Huang, A. Chaves, C. Song, V. O. Özçelik, T. Low, and H. Yan, *Nature Communications* **8**, 14071 (2017).
 - [21] J. Kang, L. Zhang, and S.-H. Wei, *The journal of physical chemistry letters* **7**, 597 (2016).
 - [22] D. A. Miller, *Quantum mechanics for scientists and engineers* (Cambridge University Press, 2008).
 - [23] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, 55 (Courier Corporation, 1964).
 - [24] M. Rudan, *Physics of Semiconductor Devices* (Springer, 2015).
 - [25] J. Bardeen, *Physical Review Letters* **6**, 57 (1961).
 - [26] W. A. Harrison, *Physical Review* **123**, 85 (1961).
 - [27] C. B. Duke, *Tunneling in solids*, vol. 10 (Academic Pr, 1969).
 - [28] R. M. Feenstra, D. Jena, and G. Gu, *Journal of Applied Physics* **111**, 043711 (2012).
 - [29] L. Britnell, R. Gorbachev, A. Geim, L. Ponomarenko, A. Mishchenko, M. Greenaway, T. Fromhold, K. Novoselov, and L. Eaves, *Nature communications* **4**, 1794 (2013).
 - [30] C. Alper, L. Lattanzio, L. De Michielis, P. Palestri, L. Selmi, and A. M. Ionescu, *Electron Devices, IEEE Transactions on* **60**, 2754 (2013).
 - [31] S. Agarwal, J. T. Teherani, J. L. Hoyt, D. A. Antoniadis, and E. Yablonovitch, *Electron Devices, IEEE Transactions on* **61**, 1599 (2014).
 - [32] A. Anwar, B. Nabet, J. Culp, and F. Castro, *Journal of applied physics* **85**, 2663 (1999).
 - [33] Y. Anugrah, M. C. Robbins, P. A. Crowell, and S. J. Koester, *Applied Physics Letters* **106**, 103108 (2015).
 - [34] F. Stern, *Physical Review B* **5**, 4891 (1972).

Supporting information: Semi-analytical model of the contact resistance in two-dimensional semiconductors

I. MODEL OF VERTICAL TUNNELING

We compute the tunneling rate in the limit of an infinite contact length and assume that the electric potential does not depend on the longitudinal position x , which implies translational invariance along x . Let $z = 0$ be the vertical position of the metal-to-semiconductor interface. We consider a simple effective mass Hamiltonian

$$\mathcal{H} = -\frac{\hbar^2}{2}\nabla \cdot \hat{m}^{-1}\nabla + E_c(z) = \mathcal{T} + E_c(z), \quad \hat{m} = \begin{pmatrix} m_x & 0 & 0 \\ 0 & m_y & 0 \\ 0 & 0 & m_z \end{pmatrix}, \quad (\text{S1})$$

where the values of the effective masses $m_{x,y,z}$ are taken to be the same in the metal and in the semiconductor and the conduction band edge profile $E_c(z)$ is modeled as a triangular barrier ($F > 0$ is the magnitude of the vertical electric field, see Fig. S1a):

$$E_c(z) = \begin{cases} V_0, & z < 0 \\ -qFz, & 0 < z < a \\ \infty, & z > a \end{cases} \quad (\text{S2})$$

Note that this is different from the Fowler-Nordheim field-emission problem [S1] because of the presence of the hard wall at $z = a$. Suppose that an electron is launched from $z < 0$ towards the interface. The electron wavefunction will be totally reflected at $z = a$, resulting in a reflection coefficient, measured as the ratio between the probability currents of reflected and incident waves, identically equal to one at all energies. Does it mean that the tunneling probability is zero? The Bardeen Transfer Hamiltonian method [S2–S4] provides a way to overcome this difficulty: the tunneling process across the barrier is thought of as a scattering event between states localized on different sides of the junction and the corresponding transition probability is computed through Oppenheimer's version of time-dependent perturbation theory [S5].

More precisely, \mathcal{H} is taken as the perturbed Hamiltonian acting in the time interval $0 < t < t_P$. For $t < 0$, the unperturbed Hamiltonian must be identified with an Hamiltonian \mathcal{H}_L that approximates well the true Hamiltonian \mathcal{H} on the metal side of the junction but whose eigenfunctions decay in the semiconductor. We take $\mathcal{H}_L = \mathcal{T} + E_{cL}(z)$ with $E_{cL}(z)$ a potential step (Fig. S1b):

$$E_{cL}(z) = \begin{cases} V_0, & z < 0 \\ 0, & z > 0 \end{cases} \quad (\text{S3})$$

For $t > t_P$, one must choose a different unperturbed Hamiltonian \mathcal{H}_R which, conversely, approximates well \mathcal{H} on the semiconductor side of the junction but whose eigenfunctions decay in the metal. We take $\mathcal{H}_R = \mathcal{T} + E_{cR}(z)$ with $E_{cR}(z)$ a triangular well (Fig. S1c):

$$E_{cR}(z) = \begin{cases} -qFz, & z < a \\ \infty, & z > a \end{cases} \quad (\text{S4})$$

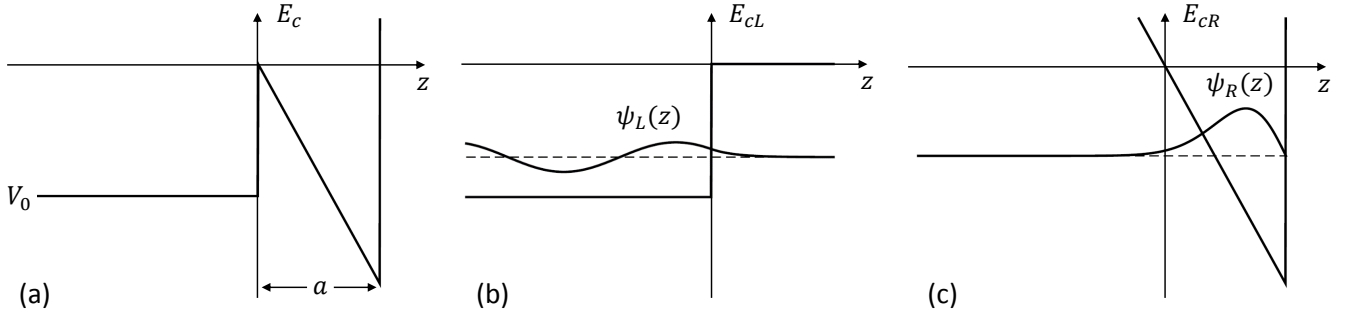


FIG. S1: Energy band profiles of the different Hamiltonians: (a) $E_c(z)$, (b) $E_{cL}(z)$, (c) $E_{cR}(z)$. The wavefunctions corresponding to the cases (b) and (c) for motion normal to the junction are also represented schematically.

It is assumed that prior to the perturbation the electron wavefunction coincides with an eigenfunction $\psi_{L,\alpha}$ of \mathcal{H}_L . Since this is not an eigenstate of \mathcal{H} , the electron wavefunction will evolve during the time interval $0 < t < t_P$ according to the time-dependent Schrödinger equation. If t_P is sufficiently large, the probability that the electron is subsequently found in the eigenstate $\psi_{R,\beta}$ of \mathcal{H}_R at $t > t_P$, is, to first order and per unit t_P ,

$$P_{\alpha\beta} = \frac{2\pi}{\hbar} |\langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle|^2 \delta(E_{L,\alpha} - E_{R,\beta}), \quad (\text{S5})$$

where $E_{L,\alpha}$ and $E_{R,\beta}$ are the eigenvalues corresponding to the initial state $\psi_{L,\alpha}$ and final state $\psi_{R,\beta}$, respectively, and δ is Dirac's delta function [S5]. The conservation of energy is related to the perturbation being constant in time. In (S5), it is assumed that each set of eigenfunctions $\psi_{L,\alpha}$ and $\psi_{R,\beta}$ is discrete and orthonormal. We consider a rectangular domain with finite sides L_x and L_y in the plane parallel to the junction and prescribe the periodic boundary conditions $\psi_{L,\alpha/R,\beta}(x=0, y, z) = \psi_{L,\alpha/R,\beta}(x=L_x, y, z)$, $\psi_{L,\alpha/R,\beta}(x, y=0, z) = \psi_{L,\alpha/R,\beta}(x, y=L_y, z)$. In addition, we consider a finite length L_z of the metal region in the z direction with the hard-wall boundary condition $\psi_{L,\alpha}(x, y, z = -L_z) = 0$. This way, both energy spectra are discrete and the corresponding eigenfunctions normalizable. Later, we will take the limit as L_x, L_y, L_z go to infinite in order to recover the continuous case. It should be noted that, contrary to the standard time-dependent perturbation theory [S6], the matrix element in (S5) must be computed between eigenstates of different Hamiltonians. Also, the perturbation Hamiltonian must be evaluated with respect to the initial Hamiltonian. The validity of (S5) rests on the assumption that the two sets of eigenstates of \mathcal{H}_L and \mathcal{H}_R are “almost orthogonal” to each other, in particular that $\langle \psi_{R,\beta} | \psi_{L,\alpha} \rangle \ll 1$ [S5]. Similarly, one has for the probability rate of the inverse transition

$$P_{\beta\alpha} = \frac{2\pi}{\hbar} |\langle \psi_{L,\alpha} | \mathcal{H} - \mathcal{H}_R | \psi_{R,\beta} \rangle|^2 \delta(E_{L,\alpha} - E_{R,\beta}) = P_{\alpha\beta}, \quad (\text{S6})$$

where the last equality follows from the delta function and the Hermiticity of the various Hamiltonians.

From (S2)-(S4) we get

$$\mathcal{H} = \begin{cases} \mathcal{H}_L, & z < 0 \\ \mathcal{H}_R, & z > 0 \end{cases} \quad (\text{S7})$$

which means that \mathcal{H} satisfies the separability property of Bardeen's model Hamiltonian, i.e., that $\mathcal{H} - \mathcal{H}_L \neq 0$ only in regions of space where $\mathcal{H} - \mathcal{H}_R \equiv 0$ [S2]¹. Since $\mathcal{H} - \mathcal{H}_L \equiv 0$ for $z < 0$, the matrix element in (S5) can be written as

$$\begin{aligned} \langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle &= \int_0^{L_x} dx \int_0^{L_y} dy \int_{-\infty}^{\infty} dz \psi_{R,\beta}^* (\mathcal{H} - \mathcal{H}_L) \psi_{L,\alpha} \\ &= \int_{\Omega_R} \psi_{R,\beta}^* (\mathcal{H} - \mathcal{H}_L) \psi_{L,\alpha} d^3r, \end{aligned} \quad (\text{S9})$$

where $\mathbf{r} = (x, y, z)$ and $\Omega_R = \{\mathbf{r}; 0 < x < L_x, 0 < y < L_y, 0 < z < \infty\}$. Noting that $\mathcal{H} - \mathcal{H}_R \equiv 0$ for $z > 0$, we can get the symmetric expression

$$\begin{aligned} \langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle &= \int_{\Omega_R} [\psi_{R,\beta}^* (\mathcal{H} - \mathcal{H}_L) \psi_{L,\alpha} - \psi_{L,\alpha} (\mathcal{H} - \mathcal{H}_R) \psi_{R,\beta}^*] d^3r \\ &= \int_{\Omega_R} [\psi_{R,\beta}^* (\mathcal{T} - E_{L,\alpha}) \psi_{L,\alpha} - \psi_{L,\alpha} (\mathcal{T} - E_{R,\beta}) \psi_{R,\beta}^*] d^3r. \end{aligned} \quad (\text{S10})$$

Due to the delta function in (S5), we are only interested in the case $E_{L,\alpha} = E_{R,\beta}$, for which

$$\langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle = \int_{\Omega_R} [\psi_{R,\beta}^* \mathcal{T} \psi_{L,\alpha} - \psi_{L,\alpha} \mathcal{T} \psi_{R,\beta}^*] d^3r. \quad (\text{S11})$$

¹ Bardeen's theory is often introduced by writing the Hamiltonian in the form $\mathcal{H} = \mathcal{H}_L + \mathcal{H}_R + \mathcal{H}_T$, with \mathcal{H}_T being the “transfer” Hamiltonian, despite the fact that Bardeen himself did not make use of such decomposition in its original paper [S2]. In our case, we get from (S7)

$$\mathcal{H}_T = \begin{cases} -\mathcal{H}_R, & z < 0 \\ -\mathcal{H}_L, & z > 0 \end{cases} \quad (\text{S8})$$

but this Hamiltonian does not correspond to either of the perturbation Hamiltonians that appear in (S5) or (S6). See also the discussion in [S4].

Applying Green's theorem, we finally get

$$\langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle = -i\hbar \int_{\Sigma_R} \mathbf{n} \cdot \mathbf{J}_{\beta\alpha} d^2r, \quad (\text{S12})$$

where Σ_R is the surface of Ω_R , \mathbf{n} is the unit vector normal to Σ_R pointing in the outward direction, and

$$\mathbf{J}_{\beta\alpha} = -\frac{i\hbar}{2} \hat{m}^{-1} [\psi_{R,\beta}^* \nabla \psi_{L,\alpha} - \psi_{L,\alpha} \nabla \psi_{R,\beta}^*] \quad (\text{S13})$$

is the matrix element of the probability current density operator between the states $\psi_{L,\alpha}$ and $\psi_{R,\beta}$.

Let us now start to pick up all the ingredients that we need to calculate (S12). The eigenfunctions and corresponding eigenvalues of \mathcal{H}_L are [S6]

$$\psi_{L,\alpha}(\mathbf{r}) \equiv \psi_L(\mathbf{k}_L; \mathbf{r}) = \frac{1}{\sqrt{L_x L_y}} e^{i(k_{xL}x + k_{yL}y)} b \times \begin{cases} 0, & z < -L_z \\ \cos(k_z z + \varphi), & -L_z < z < 0 \\ \cos(\varphi) e^{-\kappa z}, & z > 0 \end{cases} \quad (\text{S14})$$

$$E_{L,\alpha} \equiv E_L(\mathbf{k}_L) = \frac{\hbar^2}{2} \left(\frac{k_{xL}^2}{m_x} + \frac{k_{yL}^2}{m_y} \right) + E_z, \quad (\text{S15})$$

$$E_z = V_0 + \frac{\hbar^2 k_z^2}{2m_z}, \quad (\text{S16})$$

$$\kappa = \frac{\sqrt{-2m_z E_z}}{\hbar}, \quad (\text{S17})$$

$$\varphi = \arctan(\kappa/k_z), \quad (\text{S18})$$

where $\mathbf{k}_L = (k_{xL}, k_{yL}, k_z)$ and it is assumed that $E_z < 0$ ². Because of the periodic boundary conditions, the transverse components of the wavevector are quantized as $k_{xL} = 2\pi l/L_x$, $k_{yL} = 2\pi m/L_y$ (l, m integers). As for k_z , the allowed values are the roots of the transcendental equations $k_z L_z - \pi(n - 1/2) = \varphi$ in the interval $0 < k_z < \sqrt{-2m_z V_0}/\hbar$. For large L_z , we have $k_z \approx \pi n/L_z$ (n positive integer). The constant b can be obtained from the normalization condition

$$1 = |b|^2 \left[\int_{-L_z}^0 \cos^2(k_z z + \varphi) dz + \cos^2(\varphi) \int_0^\infty e^{-2\kappa z} dz \right] = \frac{|b|^2}{2} \left(L_z + \frac{1}{\kappa} \right) \approx |b|^2 \frac{L_z}{2}, \quad (\text{S19})$$

where only the leading term in L_z has been kept. Thus, up to an unimportant phase, $b = \sqrt{2/L_z}$.

The solutions of the eigenvalue problem of \mathcal{H}_R are [S7]

$$\psi_{R,\beta}(\mathbf{r}) \equiv \psi_{R,i}(\mathbf{k}_{\parallel R}; \mathbf{r}) = \frac{1}{\sqrt{L_x L_y}} e^{i(k_{xR}x + k_{yR}y)} c \times \begin{cases} \text{Ai}(\zeta), & z < a \\ 0, & z > a \end{cases} \quad (\text{S20})$$

$$E_{R,\beta} \equiv E_{R,i}(\mathbf{k}_{\parallel R}) = \frac{\hbar^2}{2} \left(\frac{k_{xR}^2}{m_x} + \frac{k_{yR}^2}{m_y} \right) + E_i, \quad (\text{S21})$$

$$E_i = -qF \left(a - \frac{|\zeta_i|}{k_F} \right), \quad (\text{S22})$$

$$\zeta = -k_F \left(z + \frac{E_i}{qF} \right), \quad (\text{S23})$$

$$k_F = \left(\frac{2m_z qF}{\hbar^2} \right)^{1/3}, \quad (\text{S24})$$

where $\mathbf{k}_{\parallel R} = (k_{xR}, k_{yR})$, Ai is Airy's function and ζ_i are its zeros, which can be approximated as (i positive integer) [S8]

$$\zeta_i \approx - \left[\frac{3\pi}{8} (4i - 1) \right]^{2/3}. \quad (\text{S25})$$

² We limit the discussion to the case $E_z < 0$ because (S5) loses validity if $\langle \psi_{R,\beta} | \psi_{L,\alpha} \rangle$ is not a small number.

The constant c can be obtained from the normalization condition

$$1 = |c|^2 \int_{-\infty}^a \text{Ai}^2(\zeta) d\zeta = \frac{|c|^2}{k_F} \int_{\zeta_i}^{\infty} \text{Ai}^2(\zeta) d\zeta, \quad (\text{S26})$$

where the last integral can be evaluated using integration by parts and the fact that Ai is a solution of Airy's equation $\text{Ai}'' = \zeta \text{Ai}$ (the prime symbol indicates derivative with respect to ζ):

$$\int_{\zeta_i}^{\infty} \text{Ai}^2(\zeta) d\zeta = - \int_{\zeta_i}^{\infty} 2\text{Ai}(\zeta) \text{Ai}'(\zeta) \zeta d\zeta = - \int_{\zeta_i}^{\infty} 2\text{Ai}'(\zeta) \text{Ai}''(\zeta) d\zeta = \text{Ai}'^2(\zeta_i). \quad (\text{S27})$$

Therefore, $c = \sqrt{k_F}/|\text{Ai}'(\zeta_i)|$, in which we can use the approximated expression [S8]

$$|\text{Ai}'(\zeta_i)| \approx \frac{1}{\sqrt{\pi}} \left[\frac{3\pi}{8} (4i-1) \right]^{1/6}. \quad (\text{S28})$$

The surface Σ_R in (S12) is made up of six faces. By inserting (S14) and (S20) into (S12), it can be shown that the integrals over the two faces at $x = 0$ and $x = L_x$, as well as the integrals over the two faces at $y = 0$ and $y = L_y$, cancel out each other exactly³. The integral over the face at $z = \infty$ is also zero because $\psi_{L,\alpha}$ is vanishingly small. We are only left with the integral over the face at $z = 0$:

$$\begin{aligned} \langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle &= \frac{\hbar^2}{2m_z} \sqrt{\frac{2}{L_z}} \frac{\sqrt{k_F}}{|\text{Ai}'(\zeta_i)|} \cos(\varphi) \left[\text{Ai}(\zeta) \frac{d}{dz} e^{-\kappa z} - e^{-\kappa z} \frac{d}{dz} \text{Ai}(\zeta) \right]_{z=0} \\ &\quad \times \frac{1}{L_x} \int_0^{L_x} e^{i(k_{xL} - k_{xR})x} dx \frac{1}{L_y} \int_0^{L_y} e^{i(k_{yL} - k_{yR})y} dy \\ &= -\frac{\hbar^2}{2m_z} \sqrt{\frac{2}{L_z}} \frac{\sqrt{k_F}}{|\text{Ai}'(\zeta_i)|} \cos(\varphi) [\kappa \text{Ai}(\zeta_0) - k_F \text{Ai}'(\zeta_0)] \delta_{k_{xL}, k_{xR}} \delta_{k_{yL}, k_{yR}}, \end{aligned} \quad (\text{S29})$$

where

$$\zeta_0 \equiv \zeta(z=0) = -k_F \frac{E_i}{qF} = k_F a - |\zeta_i| \quad (\text{S30})$$

and δ is Kronecker's delta function. Conservation of transverse momentum is a consequence of the translational symmetry along x and y . Combined with energy conservation, it implies that $E_z = E_i$ and thus $\zeta_0 = (\kappa/k_F)^2$. Assuming $\zeta_0 \gg 1$ (which is consistent with $\langle \psi_{R,\beta} | \psi_{L,\alpha} \rangle \ll 1$), we can substitute in (S29) the asymptotic expressions

$$\text{Ai}(\zeta_0) \approx \frac{e^{-\gamma_0}}{2\sqrt{\pi}\zeta_0^{1/4}}, \quad (\text{S31})$$

$$\text{Ai}'(\zeta_0) \approx -\frac{\zeta_0^{1/4} e^{-\gamma_0}}{2\sqrt{\pi}}, \quad (\text{S32})$$

where $\gamma_0 = (2/3)\zeta_0^{3/2}$ [S8], to get

$$\begin{aligned} \langle \psi_{R,\beta} | \mathcal{H} - \mathcal{H}_L | \psi_{L,\alpha} \rangle &= -\frac{\hbar^2}{2m_z} \sqrt{\frac{2}{L_z}} \frac{\sqrt{k_F}}{|\text{Ai}'(\zeta_i)|} \cos(\varphi) \left[\kappa + k_F \sqrt{\zeta_0} \right] \frac{e^{-\gamma_0}}{2\sqrt{\pi}\zeta_0^{1/4}} \delta_{k_{xL}, k_{xR}} \delta_{k_{yL}, k_{yR}} \\ &= -\frac{\hbar^2}{2m_z} \sqrt{\frac{2}{L_z}} \frac{k_F}{|\text{Ai}'(\zeta_i)|} \frac{2k_z \sqrt{\kappa}}{\sqrt{k_z^2 + \kappa^2}} \frac{e^{-\gamma_0}}{2\sqrt{\pi}} \delta_{k_{xL}, k_{xR}} \delta_{k_{yL}, k_{yR}}. \end{aligned} \quad (\text{S33})$$

Finally, plugging (S33) into (S5), we obtain

$$P_{\alpha\beta} = \frac{1}{h} \left(\frac{\hbar^2}{2m_z} \right)^2 \frac{2\pi}{L_z} \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} \frac{4k_z^2 \kappa}{k_z^2 + \kappa^2} e^{-2\gamma_0} \delta_{k_{xL}, k_{xR}} \delta_{k_{yL}, k_{yR}} \delta(E_{L,\alpha} - E_{R,\beta}). \quad (\text{S34})$$

³ For example, $J_{x,\beta\alpha}(x = L_x, y, z) - J_{x,\beta\alpha}(x = 0, y, z) \propto e^{i(k_{xL} - k_{xR})L_x} - 1 = 0$ because of the periodic boundary conditions.

II. MODEL OF LONGITUDINAL DIFFUSION

Suppose that the states in the metal (L) are populated according to a Fermi-Dirac distribution with Fermi level μ_L :

$$f_L(E_{L,\alpha}) = \frac{1}{\exp\left(\frac{E_{L,\alpha} - \mu_L}{k_B T}\right) + 1} \quad (\text{S35})$$

with k_B Boltzmann's constant and T the temperature. As for the semiconductor, we cannot assume that the states are in equilibrium because a current has to flow in the x direction as shown in Fig. 1 of the main text. In order to compute the population of such states, we assume semiclassical diffusive transport and make use of Boltzmann's transport equation [S6].

Let $f_i(x, \mathbf{k}_{\parallel R})$ be the distribution function in the four-dimensional phase space associated with the i -th subband ⁴. Under the assumption that the electric potential is uniform along x , Boltzmann's equation reads

$$v_x \frac{\partial f_i}{\partial x} = C, \quad 0 < x < L_x \quad (\text{S36})$$

where $v_x = \hbar k_{xR}/m_x$ is the longitudinal carrier velocity and the transitions from the metal to the semiconductor and vice versa due to vertical tunneling are included through a collision term C ⁵:

$$\begin{aligned} C &= \sum_{\alpha} f_L(E_{L,\alpha}) P_{\alpha\beta} (1 - f_i) - f_i P_{\beta\alpha} [1 - f_L(E_{L,\alpha})] \\ &= \sum_{\alpha} P_{\alpha\beta} [f_L(E_{L,\alpha}) - f_i]. \end{aligned} \quad (\text{S37})$$

Note that expression (S37) takes into account Pauli's exclusion principle. Using (S34), we get

$$C = [f_L(E_{R,\beta}) - f_i] \sum_{\alpha} P_{\alpha\beta} = \frac{f_L(E_{R,\beta}) - f_i}{\tau_i}, \quad (\text{S38})$$

where the relaxation time τ_i is defined as

$$\frac{1}{\tau_i} = \sum_{\alpha} P_{\alpha\beta} = \sum_{k_z} \frac{1}{h} \left(\frac{\hbar^2}{2m_z} \right)^2 \frac{2\pi}{L_z} \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} \frac{4k_z^2 \kappa}{k_z^2 + \kappa^2} e^{-2\gamma_0} \delta(E_z - E_i). \quad (\text{S39})$$

Going to the limit of large L_z , we can replace

$$\sum_{k_z} \rightarrow \int \frac{L_z}{\pi} dk_z \quad (\text{S40})$$

so that

$$\frac{1}{\tau_i} = \int_0^{\frac{\sqrt{-2m_z V_0}}{\hbar}} dk_z \frac{1}{h} \left(\frac{\hbar^2}{2m_z} \right)^2 2 \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} \frac{4k_z^2 \kappa}{k_z^2 + \kappa^2} e^{-2\gamma_0} \delta(E_z - E_i) \quad (\text{S41})$$

and, with the change of variables $k_z \rightarrow E_z$,

$$\begin{aligned} \frac{1}{\tau_i} &= \int_{V_0}^0 dE_z \frac{1}{h} \frac{\hbar^2}{2m_z} \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} \frac{4\sqrt{-E_z(E_z - V_0)}}{-V_0} e^{-2\gamma_0} \delta(E_z - E_i) \\ &= \begin{cases} \frac{1}{h} \frac{\hbar^2}{2m_z} \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} \frac{4\sqrt{-E_i(E_i - V_0)}}{-V_0} e^{-2\gamma_0}, & V_0 < E_i < 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (\text{S42})$$

⁴ f_i is independent of y because of the translational symmetry along y .

⁵ Other types of scattering, which could be responsible for a finite carrier mobility in the semiconductor, are here neglected.

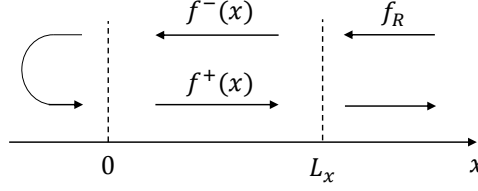


FIG. S2: Schematic description of the current fluxes and the boundary conditions of Boltzmann's equation.

Besides the subband index i , the relaxation time depends on the parameters m_z , F , a , and V_0 . It can be shown that, replacing the eigenfunctions (S14) of \mathcal{H}_L by their WKB [S9] approximation

$$\psi_{L,\alpha}^{\text{WKB}}(\mathbf{r}) = \frac{1}{\sqrt{L_x L_y}} e^{i(k_x L x + k_y L y)} \sqrt{\frac{2}{L_z}} \times \begin{cases} 0, & z < -L_z \\ \cos(k_z z + \frac{\pi}{4}), & -L_z < z < 0 \\ \frac{1}{2} \sqrt{\frac{k_z}{\kappa}} e^{-\kappa z}, & z > 0 \end{cases} \quad (\text{S43})$$

the expression of τ_i simplifies to

$$\frac{1}{\tau_i^{\text{WKB}}} = \begin{cases} \frac{1}{h} \frac{\hbar^2}{2m_z} \frac{k_F^2}{\text{Ai}'^2(\zeta_i)} e^{-2\gamma_0}, & V_0 < E_i < 0 \\ 0, & \text{otherwise} \end{cases} \quad (\text{S44})$$

where V_0 appears only as an energy cut-off. This last formulation, which does not depend on the precise bandstructure of the metal, could be useful for treating injection from the metal to the valence band of the semiconductor.

Let f_i^+ and f_i^- denote the distribution functions of right-going and left-going states, respectively, i.e., $f_i^\pm(x, k_{xR}, k_{yR}) = f_i(x, \pm k_{xR}, k_{yR})$ with $k_{xR} > 0$. We impose the boundary conditions

$$f_i^-(x = L_x, \mathbf{k}_{\parallel R}) = f_R(E_{R,\beta}), \quad (\text{S45})$$

$$f_i^+(x = 0, \mathbf{k}_{\parallel R}) = f_i^-(x = 0, \mathbf{k}_{\parallel R}), \quad (\text{S46})$$

where f_R is a Fermi-Dirac function similar to (S35) with μ_L replaced by μ_R (see Fig. S2). The latter condition makes sure that the longitudinal current vanishes at $x = 0$. The Boltzmann equations (S36) for different wavevectors are independent of each other except for $\pm k_{xR}$. The solutions are

$$f_i^\pm = -[f_L(E_{R,\beta}) - f_R(E_{R,\beta})] e^{-\frac{L_x \pm x}{|v_x| \tau_i}} + f_L(E_{R,\beta}). \quad (\text{S47})$$

The current (per unit width) from the semiconductor to the metal can be obtained by summing the net flux $|v_x| (f_i^+ - f_i^-)$ at $x = L_x$ over all semiconductor states per unit area, multiplying by 2 for spin degeneracy and multiplying by the electronic charge q :

$$\begin{aligned} I &= \frac{2q}{L_x L_y} \sum_{k_{xR} > 0} \sum_{k_{yR}} \sum_i |v_x| (f_i^+ - f_i^-)_{x=L_x} \\ &= \frac{2q}{L_x L_y} \sum_{k_{xR} > 0} \sum_{k_{yR}} \sum_i |v_x| \left(1 - e^{-\frac{2L_x}{|v_x| \tau_i}}\right) [f_L(E_{R,\beta}) - f_R(E_{R,\beta})]. \end{aligned} \quad (\text{S48})$$

Going to the limit of large L_x, L_y , we can replace

$$\sum_{k_{xR}} \rightarrow \int \frac{L_x}{2\pi} dk_{xR}, \quad \sum_{k_{yR}} \rightarrow \int \frac{L_y}{2\pi} dk_{yR} \quad (\text{S49})$$

to get

$$I = \frac{2q}{(2\pi)^2} \int_0^\infty dk_{xR} \int_{-\infty}^\infty dk_{yR} \sum_i |v_x| \left(1 - e^{-\frac{2L_x}{|v_x| \tau_i}}\right) [f_L(E_{R,\beta}) - f_R(E_{R,\beta})]. \quad (\text{S50})$$

Finally, with the change of variables $k_{xR} \rightarrow \varepsilon = \hbar^2 k_{xR}^2 / (2m_x) + E_i$, we obtain the Landauer formula [S10]

$$\begin{aligned} I &= \frac{2q}{h} \sum_i \int_{E_i}^{\infty} d\varepsilon \left(1 - e^{-\frac{2L_x}{|v_x|\tau_i}} \right) \frac{1}{2\pi} \int_{-\infty}^{\infty} dk_{yR} \left[f_L \left(\frac{\hbar^2 k_{yR}^2}{2m_y} + \varepsilon \right) - f_R \left(\frac{\hbar^2 k_{yR}^2}{2m_y} + \varepsilon \right) \right] \\ &= \frac{2q}{h} \sum_i \int_{E_i}^{\infty} d\varepsilon \left(1 - e^{-\frac{2L_x}{|v_x|\tau_i}} \right) [F_L(\varepsilon) - F_R(\varepsilon)] \\ &= \frac{2q}{h} \int_{-\infty}^{\infty} d\varepsilon \bar{T}(\varepsilon) [F_L(\varepsilon) - F_R(\varepsilon)], \end{aligned} \quad (\text{S51})$$

where the longitudinal velocity must be computed as $|v_x| = \sqrt{2(\varepsilon - E_i)/m_x}$, the supply function $F_{L/R}$ is defined as

$$F_{L/R}(\varepsilon) = \sqrt{\frac{m_y k_B T}{2\pi \hbar^2}} \mathcal{F}_{-1/2} \left(\frac{\mu_{L/R} - \varepsilon}{k_B T} \right) \quad (\text{S52})$$

with $\mathcal{F}_{-1/2}$ the Fermi-Dirac integral of order $-1/2$ [S6], \bar{T} is the transmission function

$$\bar{T}(\varepsilon) = \sum_i T_i(\varepsilon), \quad (\text{S53})$$

with the transmission probability T_i given by

$$T_i(\varepsilon) = \begin{cases} 0, & \varepsilon < E_i \\ 1 - e^{-\frac{2L_x}{\lambda_i}}, & \varepsilon > E_i \end{cases} \quad (\text{S54})$$

and $\lambda_i = |v_x|\tau_i$. For $\varepsilon > E_i$, we can have the asymptotic behaviors

$$L_x \gg \lambda_i: \quad T_i(\varepsilon) \approx 1 \quad (\text{S55})$$

$$L_x \ll \lambda_i: \quad T_i(\varepsilon) \approx \frac{2L_x}{\lambda_i} \quad (\text{S56})$$

Approximation (S55) holds, in particular, as $\varepsilon \rightarrow E_i^+$ (resonant tunneling). Note also that, when (S56) is satisfied, T_i decays as a function of energy as $1/\sqrt{\varepsilon}$.

The low bias conductance per unit width G can be evaluated from (S51). Let $\mu_L = \mu$ and $\mu_R = \mu - qV_D$. We have

$$G = \left. \frac{\partial I}{\partial V_D} \right|_{V_D=0} = \frac{2q^2}{h} \int_{-\infty}^{\infty} d\varepsilon \bar{T}(\varepsilon) \left(-\frac{\partial F_0}{\partial \varepsilon} \right), \quad (\text{S57})$$

where $F_0(\varepsilon)$ is given by (S52) with $\mu_{L/R}$ replaced by μ .

Assume for simplicity that only the first subband contributes to transport. In the two limiting cases when either (S55) or (S56) are satisfied over the whole energy range of interest for transport (from E_1 to few $k_B T$'s above $\max\{E_1, \mu\}$), it is possible to derive analytical expressions for G . If $L_x \gg \lambda_1$ in this energy range, it follows immediately from (S57)

$$G = \frac{2q^2}{h} F_0(E_1). \quad (\text{S58})$$

To work out the expression of G in the other limiting case when $L_x \ll \lambda_1$ in most of the energy window for transport⁶, it is convenient to go back to the double-integral formulation of the tunneling current in (S50) and do the change of variables $k_{xR} \rightarrow E_{R,\beta}$:

$$I = \frac{2q}{h} \int_{E_1}^{\infty} dE_{R,\beta} \left(\frac{1}{\pi} \int_0^{k_{y\max}} dk_{yR} \frac{2L_x}{|v_x|\tau_1} \right) [f_L(E_{R,\beta}) - f_R(E_{R,\beta})], \quad (\text{S59})$$

⁶ The inequality does not hold for energies close to E_1 but their contribution becomes increasingly smaller as τ_1 increases.

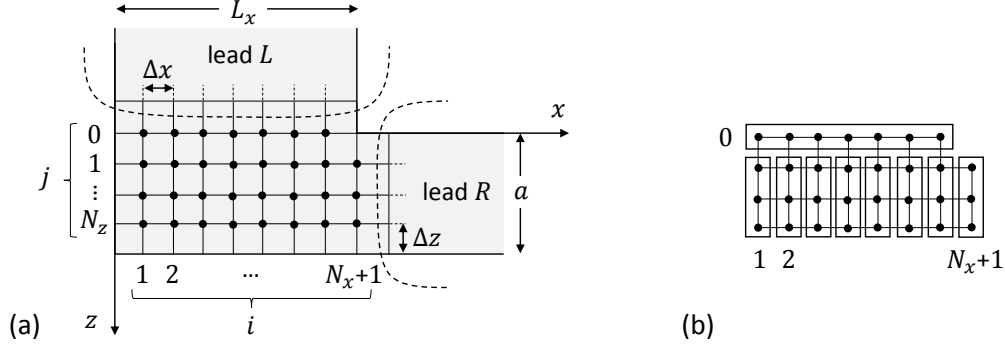


FIG. S3: (a) Simulation domain and rectangular grid of the Green function calculation. The discretization steps are $\Delta x = 0.3$ nm and $\Delta z = 0.2$ nm. The black dots indicate the grid nodes that compose the “channel” region. (b) Partitioning of the channel region into layers for the Green function algorithm.

where $k_{y\max} = \sqrt{2m_y(E_{R,\beta} - E_1)}/\hbar$ and $|v_x| = \sqrt{2[E_{R,\beta} - \hbar^2 k_{yR}^2/(2m_y) - E_1]/m_x}$. The integral over the transverse wavevector can be easily computed with the change of variables $k_{yR} \rightarrow \arcsin(k_{yR}/k_{y\max})$ to give

$$\begin{aligned} I &= \frac{2q}{h} \int_{E_1}^{\infty} dE_{R,\beta} \frac{k_{y\max} L_x}{\sqrt{2(E_{R,\beta} - E_1)/m_x} \tau_1} [f_L(E_{R,\beta}) - f_R(E_{R,\beta})] \\ &= \frac{2q}{h} \frac{\sqrt{m_x m_y}}{\hbar} \frac{L_x}{\tau_1} \int_{E_1}^{\infty} dE_{R,\beta} [f_L(E_{R,\beta}) - f_R(E_{R,\beta})]. \end{aligned} \quad (\text{S60})$$

Letting $\mu_L = \mu$ and $\mu_R = \mu - qV_D$, we finally get

$$G = \left. \frac{\partial I}{\partial V_D} \right|_{V_D=0} = \frac{2q^2}{h} \frac{\sqrt{m_x m_y}}{\hbar} \frac{L_x}{\tau_1} \int_{E_1}^{\infty} dE_{R,\beta} \left(-\frac{\partial f_0}{\partial E_{R,\beta}} \right) = \frac{2q^2}{h} \frac{\sqrt{m_x m_y}}{\hbar} \frac{L_x}{\tau_1} f_0(E_1), \quad (\text{S61})$$

where f_0 is a Fermi-Dirac function similar to (S35) with μ_L replaced by μ .

III. GREEN FUNCTION ALGORITHM

We discretize the Hamiltonian in (S1) using finite differences on a two-dimensional rectangular grid (Fig. S3a). The same linear potential profile as in (S2) is assumed. The transmission function is computed as [S10]

$$\overline{T}(\varepsilon) = \text{tr} [\Gamma^L G^r \Gamma^R G^a], \quad (\text{S62})$$

where the symbol tr indicates the trace, G^r is the retarded Green function, $G^a = G^{r\dagger}$, $\Gamma^{L/R} = i(\Sigma^{r,L/R} - \Sigma^{a,L/R})$, with $\Sigma^{r,L/R}$ the retarded self-energy representing the renormalization of the Hamiltonian of the channel region (black dots in Fig. S3a) due to the presence of the semi-infinite left/right lead, and $\Sigma^{a,L/R} = (\Sigma^{r,L/R})^\dagger$. The channel region is partitioned into layers as shown in Fig. S3b. Using matrix block notation and noting that the only non-null block of $\Sigma^{r,L}$ is $\Sigma_{0,0}^{r,L}$ and the only non-null block of $\Sigma^{r,R}$ is $\Sigma_{N_x+1,N_x+1}^{r,R}$, (S62) can be rewritten as

$$\overline{T}(\varepsilon) = \text{tr} [\Gamma_{0,0}^L G_{0,N_x+1}^r \Gamma_{N_x+1,N_x+1}^R G_{N_x+1,0}^a]. \quad (\text{S63})$$

The self-energy of the left lead is computed analytically using the prescription given in [S10]:

$$\Sigma_{0,0}^{r,L}(i,i') = \sum_{m=1}^{N_x} \chi_m(i) \sigma_m \chi_m(i'), \quad (\text{S64})$$

$$\chi_m(i) = \sqrt{\frac{2}{N_x + 1}} \sin(k_x i), \quad (\text{S65})$$

$$\sigma_m = t_z \times \begin{cases} \lambda - 1 + \sqrt{\lambda^2 - 2\lambda}, & \lambda < 0 \\ \lambda - 1 - \sqrt{\lambda^2 - 2\lambda}, & \lambda > 2 \\ \lambda - 1 - i\sqrt{2\lambda - \lambda^2}, & 0 < \lambda < 2 \end{cases} \quad (\text{S66})$$

$$\lambda = \frac{\varepsilon - V_0 - 2t_x(1 - \cos k_x)}{2t_z}, \quad (\text{S67})$$

$$k_x = \frac{\pi m}{N_x + 1}, \quad (\text{S68})$$

where $t_x = \hbar^2/(2m_x \Delta_x^2)$ and similarly for t_z . The self-energy of the right lead is obtained numerically using a well-known iterative algorithm [S11]. The matrix block G_{0,N_x+1}^r is computed through a combination of the recursive and decimation algorithms [S12], modified so as to treat a non-tridiagonal-block Hamiltonian matrix. Let $A = \varepsilon I - H_C - \Sigma^{r,L} - \Sigma^{r,R}$, where H_C is the Hamiltonian matrix of the channel region alone, and define $\delta_1^{(0)} = A_{0,0}$, $\delta_2^{(0)} = A_{1,1}$, $\alpha^{(0)} = A_{0,1}$, $\beta^{(0)} = A_{1,0}$. The algorithm consists in eliminating the layers from 1 to N_x with the formulas

$$\begin{aligned} \delta_1^{(n)} &= \delta_1^{(n-1)} - \alpha^{(n-1)} \left[\delta_2^{(n-1)} \right]^{-1} \beta^{(n-1)}, \\ \delta_2^{(n)} &= A_{n+1,n+1} - A_{n+1,n} \left[\delta_2^{(n-1)} \right]^{-1} A_{n,n+1}, \\ \alpha^{(n)} &= -\alpha^{(n-1)} \left[\delta_2^{(n-1)} \right]^{-1} A_{n,n+1} + A_{0,n+1}, \\ \beta^{(n)} &= -A_{n+1,n} \left[\delta_2^{(n-1)} \right]^{-1} \beta^{(n-1)} + A_{n+1,0} \end{aligned} \quad (\text{S69})$$

for $n = 1, \dots, N_x$, where it is understood that $A_{0,N_x+1} = A_{N_x+1,0}^\dagger = 0$. At the end, the required matrix block of the Green function can be obtained as

$$G_{0,N_x+1}^r = - \left[\delta_1^{(N_x)} \right]^{-1} \alpha^{(N_x)} \left\{ \delta_2^{(N_x)} - \beta^{(N_x)} \left[\delta_1^{(N_x)} \right]^{-1} \alpha^{(N_x)} \right\}^{-1}. \quad (\text{S70})$$

-
- [S1] R. H. Fowler and L. Nordheim, in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (The Royal Society, 1928), vol. 119, pp. 173–181.
[S2] J. Bardeen, *Physical Review Letters* **6**, 57 (1961).
[S3] W. A. Harrison, *Physical Review* **123**, 85 (1961).
[S4] C. B. Duke, *Tunneling in solids*, vol. 10 (Academic Pr, 1969).
[S5] J. R. Oppenheimer, *Physical review* **31**, 66 (1928).
[S6] M. Rudan, *Physics of Semiconductor Devices* (Springer, 2015).
[S7] D. A. Miller, *Quantum mechanics for scientists and engineers* (Cambridge University Press, 2008).
[S8] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, 55 (Courier Corporation, 1964).
[S9] A. Messiah, *Quantum mechanics, vol. 1* (North-Holland, Amsterdam, 1961).
[S10] S. Datta, *Electronic transport in mesoscopic systems* (Cambridge university press, 1997).
[S11] M. L. Sancho, J. L. Sancho, J. L. Sancho, and J. Rubio, *Journal of Physics F: Metal Physics* **15**, 851 (1985).
[S12] T. Low and J. Appenzeller, *Physical Review B* **80**, 155406 (2009).